

# Using Automated Rhyme Detection to Characterize Rhyming Style in Rap Music

HUSSEIN HIRJEE

*Cheriton School of Computer Science, University of Waterloo*

DANIEL G. BROWN

*Cheriton School of Computer Science, University of Waterloo*

**ABSTRACT:** Imperfect and internal rhymes are two important features in rap music previously ignored in the music information retrieval literature. We developed a method of scoring potential rhymes using a probabilistic model based on phoneme frequencies in rap lyrics. We used this scoring scheme to automatically identify internal and line-final rhymes in song lyrics and demonstrated the performance of this method compared to rules-based models. We then calculated higher-level rhyme features and used them to compare rhyming styles in song lyrics from different genres, and for different rap artists. We found that these detected features corresponded to real-world descriptions of rhyming style and were strongly characteristic of different rappers, resulting in potential applications to style-based comparison, music recommendation, and authorship identification.

Submitted 2010 May 26; accepted 2010 August 29.

**KEYWORDS:** *song lyrics, phonetic similarity, rhyme, hip hop, artist classification*

SONG lyrics have received relatively little attention in music information retrieval, but can provide data about song style or content that is missing from raw audio files or user-input tags. Recent work focusing on lyrics (Fujihara, 2008; Kleedorfer, 2008; Wei, 2007) uses the meaning of lyric text words to extract song topic, theme, or mood information; the pattern and sound of the words themselves is usually ignored.

These sound features are central to rap music, providing information about vocal delivery and rhyme scheme. These data can be characteristic of different rappers, as MCs often boast of the uniqueness and superiority of their rhyming style (Bradley, 2009). Lyric rhymes have previously been studied as an aid in predicting musical genres (Mayer, 2008), but this prior work ignores two stylistic features of rap lyrics: imperfect rhymes, where syllable end sounds are similar, but not identical; and internal rhyme, which occurs in the middle of lines.

To study these features, we developed a system for automatic detection of rap music rhymes. We trained a probabilistic scoring model of rhymes using a corpus of rap lyrics known to be rhyming, using ideas derived from bioinformatics. We then used this model to find and categorize various rhymes in different song lyrics, and assessed the model's success. High-level statistical rhyme scheme features we calculated allowed us to quantitatively model and compare rhyming styles between artists and genres. These features correlated with real-world notions of rapping style and we identified trends in their use in hip hop music over time. Finally, we used these rhyme features to classify rappers and investigated potential applications of rhyme stylometry. This article is the expanded version of a conference paper (Hirjee, 2009) presented at ISMIR 2009.

## BACKGROUND

Hip hop music is characterized by lyrics with intermittent rhymes being rhythmically chanted (rapped) to an accompanying beat. In "Old School" rap (from the late 1970s to mid 1980s), lyrics typically followed a simple pattern and contained a single rhyme falling on the fourth beat of each bar (Bradley, 2009). Contemporary rap features more varied delivery and many complex rhyme elements that are often overlooked. Key among these are rhymes that are imperfect, extended, or internal.

## Imperfect Rhymes

Holtman (1996) provides a good overview of the abundance of imperfect rhyme (also called slant rhyme) in rap lyrics and identifies some examples of their use in Eric B. and Rakim's *Let The Rhythm Hit 'Em* (1990). A normal rhyme involves two syllables that share the same nucleus (vowel) and coda (ending consonants). Two syllables form an imperfect rhyme if one of these two parts does not correspond exactly. However, she argues that these types of rhymes are not just composed of vowels and consonants being paired randomly: there is a limit to the amount of dissimilarity in these rhymes, determined by the shared articulatory features of matching phonemes.

In Holtman's hierarchy, the most similar consonants are nasals, fricatives, and plosives differing only in place of articulation, as in the line-ending /m/ and /n/ phonemes in:

Entertain and tear you out of your **frame**  
Leave you in a puddle of blood, then let it **rain**, (Eric B. & Rakim, 1990)

as well as the /k/-/t/ from "black"- "fat" and /t/-/p/ from "coat"- "rope" pairs in:

Cool, I heat you up like a **black mink coat**  
Hug your neck like a **fat gold rope**. (Eric B. & Rakim, 1990)

(Rhyming syllables in quoted lyrics are displayed with the same font style.)

Less similar consonant pairs include those with the same place of articulation, but differing in voice or continuancy, such as the /k/ and /g/ pair in:

Bring a bullet-proof vest, nothin' to **ricochet**  
Ready to aim at the brain, now what the **trigger say?** (Eric B. & Rakim, 1990)

Though vowel identity tends to be preserved in rhymes, nonidentical vowels are most similar when differing only in height or "length" (advanced tongue root), such as the penultimate vowels (/ɛ/ and /eɪ/) in:

I'm the alpha, with no **omega**  
Beginning without the, end so **play the**, (Eric B. & Rakim, 1990)

or the /a/ and /ɔ/ in:

Beats and bullets pass me, none on **target**  
They want the R hit, but watch the **god get**. (Eric B. & Rakim, 1990)

Less similar vowel pairs differ in front/back position such as the /ɛ/ and /ɔ/ in:

Vocabs is endless, vocals **exist**  
Rhyme goes on, so no one can **stop this**. (Eric, B & Rakim, 1990)

Holtman's work is largely taxonomic and describes known rhymes, rather than discovering them. We used a statistical model of phonetic similarity based on frequencies in actual rap lyrics to quantify these varying amounts of perceived similarity. The patterns we automatically discovered largely validate her taxonomy.

## Polysyllabic Rhymes

Rap music often features three-syllable or longer rhymes with unstressed syllables following the initial stressed pair. Also known as multisyllabic rhymes or multis, these may span multiple words, in which case they are called mosaic rhymes. Longer rhymes can also include more than one pair of stressed syllables:

Maybe my sense of **húmor gets into you**  
But girl, they can make a **perfúme from the scént of you**. (Fabolous, 2001)

(Here the accents mark the syllables with primary stress in the six-syllable rhyme.)

## Internal Rhymes

Finally, contemporary rap music features dazzlingly complex internal rhyme. Alim (2003) analyzes Pharoahe Monch's album *Internal Affairs* (1999) as a case study, and identifies chain rhymes, compound rhymes, and bridge rhymes. Chain rhymes are consecutive words or phrases in which each rhymes with the previous, as in:

New York **City gritty committee pity** the fool that  
Act **shitty** in the midst of the calm the **witty**, (Monch, 1999)

where “city”, “gritty”, “committee”, and “pity” participate in a chain since they all rhyme and follow each other contiguously.

Compound rhymes are formed when two pairs of line internal rhymes overlap within a single line. A good example of this is given in “Official”:

Yo, I stick around like hockey, now what the **puck**  
Cooler than **fuck**, **maneuver** like **Vancouver Canucks**, (Monch, 1999)

where “maneuver” and “Vancouver” are found between “fuck” and “Canucks.”

Bridge rhymes are internal rhymes spanning two lines:

**How I made it you salivated** over my **calibrated**  
RAPS that **validated** my ghetto *credibility*  
*Still I be* PACKin *agilities unseen*  
Forreal-a my killin *abilities unclean facilities*. (Monch, 1999)

Here, we called pairs in which both members are internal (such as “agilities” / “abilities”) bridge rhymes, and those where the first word or phrase is line-final (such as “calibrated” / “validated”), link rhymes.

## FINDING RHYMES AUTOMATICALLY: A PROBABILISTIC MODEL

We used a model inspired by protein homology detection techniques from bioinformatics, in which proteins are identified as sequences of amino acids (Durbin, Eddy, Krogh, & Mitchison, 1999). In this framework, a pair of proteins is modeled as two sequences of amino acid symbols generated either randomly or based on shared ancestry (homology). Using the BLOSUM (BLOCKS of amino acid SUBstitution Matrix) local alignment scoring scheme (Henikoff & Henikoff, 1992), pairs of amino acids are assigned log-odds scores based on the likelihood of their being matched in alignments of homologous proteins. A positive score indicates the pair more likely co-occurs in proteins evolved from a shared ancestor, while a negative score indicates the pair is more likely to co-occur due to chance. In a BLOSUM matrix  $M$ , the score for any two amino acids  $i$  and  $j$ , is calculated as

$$M[i, j] = \log_2 (\Pr[i, j | H] / \Pr[i, j | R]),$$

where  $\Pr[i, j | H]$  is the likelihood of  $i$  being matched to  $j$  in an alignment of two homologous proteins, while  $\Pr[i, j | R]$  is the likelihood of them being matched by chance.

These likelihoods are calculated using frequencies of amino acid pairings in alignments of proteins known to be homologous. Given an amino acid pair frequency table  $F$ , where  $F_{i,j}$  is the number of times  $i$  is matched to  $j$  in a collection of homologous protein alignments, the homology likelihood is calculated as

$$\Pr[i, j | H] = F_{i,j} / \sum_m \sum_n F_{m,n}.$$

This corresponds to the proportion of amino acid pairs in which  $i$  matches with  $j$ . The match by chance likelihood is calculated as

$$\Pr[i, j | R] = F_i \times F_j / (\sum_m F_m \times \sum_m F_m),$$

where  $F_i$  is the total number of times amino acid  $i$  appears in the collection. This is simply the product of the background frequencies of each amino acid in the pair. If a pair of protein sequences contains regions in which the amino acids align to give high scores, the pair is considered to be homologous.

In our work, we transcribed song lyrics into sets of sequences of syllables, with each sequence corresponding to a line of text. Similar to Kawahara's (2007) treatment of consonants in Japanese rap

lyrics, we used probabilistic methods to calculate similarity scores for any given pair of syllables. Our method assigned positive scores to phonemes that matched with each other in rhyming phrases more often than expected by chance, and negative scores to those that matched less often than expected by chance. Regions with syllables that, when matched to each other, had a total score surpassing a threshold were identified as rhymes.

## RHYMING SYLLABLES

To generate models of rhyming and randomly co-occurring syllables in rap lyrics, we required a data set of known rhymes. This data set corresponded to the corpus of alignments of homologous proteins used to train the BLOSUM matrix. Our training corpus included the lyrics of 31 influential albums from the “Golden Age” of rap (1984-1994), chosen because they received the highest rating of Five Mics from *The Source*, the top-selling US rap music magazine of the time (Ogbar, 1999), plus nine additional albums by influential artists from the time period (Run-DMC, LL Cool J, the Beastie Boys, Public Enemy, Eric B. and Rakim). We downloaded lyrics from the Web and manually corrected them to fix typographical errors and ensure that pairs of consecutive lines ended with matching rhymes, splitting some lines in half and repeating others when necessary. This yielded 27,956 lines of lyrics (13,978 rhymed pairs), approximately 700 lines per album.

We first transcribed plain text lyrics into sequences of phonemes using a wrapper we built around the Carnegie Mellon University (CMU) Pronouncing Dictionary (Lenzo, 2007), which gives phonetic transcriptions for over 100,000 words in North American English. The transcriptions contain 39 phonemes, consisting of 24 consonants, including affricates such as /tʃ/ and /dʒ/, and 15 vowels, including diphthongs like /aʊ/ and /ɔi/ (IPA, 1999). The vowels include metrical stress markings indicating whether they receive primary (1), secondary (2), or no stress (0). Thus, for each word in the dictionary, the transcription provides the speech sounds (phonemes), as well as the prosody (the pattern of emphasis placed on each syllable when pronounced.) To avoid the complications and computational complexity required to evaluate all possible transcriptions for heteronyms and other words with numerous pronunciations, we selected the first transcription for each word, corresponding to the most common pronunciation.

We augmented the dictionary with common elements of hip hop vernacular and slang, including terms such as “DJ,” “basehead,” and “AK-47,” as well as a wide variety of profanity. To accommodate for variations in spelling and pronunciation in the lyrics, we implemented rules to transform pronunciations for common occurrences of these variations, such as the “-in” ending in “runnin’,” or the “-a” ending in “brotha” or “killa.” Finally, we reduced the stress assigned to about 30 common one-syllable words of minor significance in rhyme (“a,” “I,” “and,” etc.) to better model their actual realizations in rap performance. To handle words not found in the augmented dictionary, we added the Naval Research Laboratory’s text-to-phoneme rules (Elovitz, Johnson, McHugh, & Shore, 1976). These rules provide a phonetic substitution approximating the correct pronunciation for each of the 26 letters of the alphabet, based on the characters surrounding them in the word.

## SCORING POTENTIAL RHYMES

To generate a log-odds scoring matrix for rhyming syllables, we required models for random syllables and for rhymes. For any pair of syllables  $i$  and  $j$ , the random model,  $\Pr[i,j|\text{Random}]$ , gives the likelihood of  $i$  and  $j$  being matched together by chance while the rhyme model,  $\Pr[i,j|\text{Rhyme}]$ , gives the likelihood of  $i$  and  $j$  being paired in a true rhyme. As in BLOSUM (Henikoff & Henikoff, 1992), the log-odds score was calculated as

$$M[i, j] = \ln(\Pr[i, j | \text{Rhyme}] / \Pr[i, j | \text{Random}]) .$$

Here, we assumed that rhymes included no skipped syllables, though this is not always true in rap. To avoid overfitting, we reduced each syllable to its vowel (nucleus), end consonants (coda), and stress—the relevant features for determining rhyme. We approximated the coda by taking the first half (rounded up) of the consonants between adjacent pairs of vowels. Both models were trained using the occurrence frequencies of phonemes in the training data.

In the random model, the likelihood of vowel  $a$  matching with vowel  $b$  was calculated by taking the product of the frequencies of  $a$  and  $b$ :

$$\Pr[a, b | \textit{Random}] = F_a \times F_b / (\sum_m F_m \times \sum_m F_m),$$

where  $F_a$  is the total number of times phoneme  $a$  appeared in the lyrics. The likelihoods for consonants and varying stress were calculated independently in the same manner.

For the rhyming model, the likelihood of vowels  $a$  and  $b$  being matched was calculated by taking the number of times  $a$  and  $b$  were seen matching in known rhymes, and dividing by the total number of matched vowel pairs in known rhymes. This was calculated as

$$\Pr[a, b | \textit{Rhyme}] = F_{a,b} / \sum_m \sum_n F_{m,n},$$

where  $F_{a,b}$  is the number of times vowels  $a$  and  $b$  appeared matched in the known rhymes. Then the log-odds score for the vowels was calculated as

$$\ln(\Pr[a, b | \textit{Rhyme}] / \Pr[a, b | \textit{Random}]).$$

The likelihood for consonants was more complicated since we needed to also consider unmatched consonants when aligning syllable codas of differing size. For example, the following pair from Public Enemy’s “Black Steel in the Hour of Chaos” has line final consonant clusters of /ld/ and /d/:

Cold holdin’ the load, the burden breakin’ the mold  
I ain’t lyin’ denyin’, because they’re checkin’ my code. (Public Enemy, 1988)

We used an iterated approach to match the most likely consonants. In the first pass over the training data, we considered rhymes in paired lines to be all syllables following the final primary-stressed syllable, after Holtman (1996), and produced an initial scoring matrix  $M'$  by calculating the above statistics. We aligned consonants sequentially from left to right. For the example given, the IPA transcription ends with

(the burden breakin’ the mold)  
ð ə b 'ɜː d ə n b r 'eɪ k ɪ n ð ə m 'oʊ l d  
b ɪ k 'ɔ z ð ɛ r tʃ 'ɛ k ɪ n m aɪ k 'oʊ d  
(because they’re checkin’ my code),

so the rhyme would start at the /oʊ/ vowels. The /l/ from “mold” was matched with the /d/ from “code” and the /d/ in mold was unmatched. Here, we introduced symbols /\_\*/ and /\*\_/ that we treated as consonants to allow for different penalties for different unmatched consonants at the beginning and end of codas. This distinction was useful since some consonants (such as /l/ and /r/) were more likely to be unmatched at the beginning of clusters, and others (often coronals, such as /d/ and /z/) were more likely to be unmatched at the ends of clusters. A simple example of this is found in the occurrences of “harm,” “unarmed,” and “alarmed” rhyming with “bomb” in Public Enemy’s “Louder Than A Bomb” (Public Enemy, 1988); in these cases, the words still form imperfect rhymes, despite the unmatched consonants.

In the second pass, we processed rhymes by moving backwards from the end of the line and using the initial scoring matrix  $M'$  (derived in the first pass) to calculate scores for matching syllables. We stopped when we encountered a negative score for a stressed syllable pair, and identified the start of the rhyme as the last positive-scored stressed syllable pair encountered. For the example above, the rhyme was identified as “breakin’ the mold” with “checkin’ my code.” We used  $M'$  to perform global alignment (Durbin et al., 1999) on matched codas to determine frequencies for consonants pairing with other consonants, and being unmatched at the start or end of the coda. When the codas for “mold” and “code” were aligned in this way, the /d/s matched together and the /l/ was treated as an unmatched phoneme at the start of the consonant cluster:

(the burden breakin’ the mold)  
ð ə b 'ɜː d ə n b r 'eɪ k ɪ n ð ə m 'oʊ l d  
b ɪ k 'ɔ z ð ɛ r tʃ 'ɛ k ɪ n m aɪ k 'oʊ \_ d.  
(because they’re checkin’ my code)

These updated alignments gave us a second frequency table from which we produced the rhyming model and log-odds scores for consonants and stress in the same way as for vowels. Finally, we normalized the consonant score by dividing by the length of the coda to avoid the problem of syllables with long codas having the consonant score dominate. Intuitively, “win” and “gin” rhyme as well as “splints” and “mints.” Since all of the constituent scores were log-odds, they could be added together to form a combined

probabilistic log-odds score. This form of combination assumes that all sound features are independent. This is not necessarily correct (for example, after different vowels, there are different distributions of consonants), but works well as an approximation. The final score for two given syllables  $i$  and  $j$  is the sum of the vowel score, normalized consonant score, and stress score:

$$Score(i, j) = M[i_v, j_v] + align(i_c, v_c) + M[i_{stress}, j_{stress}],$$

where  $i_v$  is the vowel in  $i$ ,  $align(i_c, v_c)$  is the score for the global alignment of the end consonants of  $i$  and  $j$ , and  $i_{stress}$  is the metrical stress markings of  $i$ . Since we only considered the pronunciations, homophones and identical words were treated in the same manner, and both generally received high scores.

Tables 1 and 2 show the pairwise scoring matrices for stressed vowels and consonants. The symbols “\_\*” and “\*\_” indicate scores for unmatched consonants at the beginning and end of codas, respectively. The similarity of vowels differing in height only appears for back vowels: /a/, /ɔ/, and /ʊ/, receiving high scores when paired. We see this less for front vowels /eɪ/, /ɛ/, /ɪ/, and /i/, though these tend not to score as negatively as other vowels when paired. In the consonant matrix, high scores for fricative pairs like (/f/, /θ/) and (/v/, /ð/), nasals (/m/, /n/), as well as plosives such as (/k/, /p/) and (/p/, /t/) largely validate Holtman’s hierarchy (Holtman, 1996). We also see an interesting effect where affricates score highly with their constituent fricatives: (/tʃ/, /ʃ/) and (/dʒ/, /ʒ/). The consonants most likely to appear unmatched at the ends of codas include /d/, /z/, /w/, /t/, and /s/, which for the most part probably correspond to common endings for verb past tenses and noun plurals, such as “trap” rhyming with “capped” or “hot” rhyming with “rocks.” (We note that “rhymes with” is not a categorical relationship in this domain: while it is not obvious if a particular listener would say “trap” rhymes with “capped,” most native English speakers would likely agree that these rhyme better than “trap” with “fit.”) The consonants most likely to be unmatched at the start of codas are the approximants /r/ and /l/, and /s/, as in “master” rhyming with “stature.”

	/a/	/æ/	/ʌ/	/ɔ/	/aʊ/	/ai/	/ɛ/	/ɜ/	/eɪ/	/ɪ/	/i/	/oʊ/	/ɔi/	/ʊ/	/u/
/a/	2.3	-3.2	-0.8	1.6	-1.7	-2.7	-7.2	-0.6	-3.9	-4.8	-3.9	-1.0	-1.7	-3.3	-3.9
/æ/		2.1	-1.5	-6.6	-1.9	-3.3	-1.5	-3.4	-1.8	-2.0	-4.3	-4.6	-4.5	-3.7	-6.7
/ʌ/			2.2	-1.2	-1.4	-1.4	-0.6	-0.2	-1.7	-0.3	-3.0	-1.0	-0.6	-0.9	-1.5
/ɔ/				3.1	-1.0	-3.8	-6.5	-1.1	-3.9	-4.2	-6.3	-0.3	-0.4	1.1	-3.3
/aʊ/					3.8	-0.3	-6.0	-4.2	-5.7	-6.0	-5.7	-2.0	-2.9	-4.5	-1.4
/ai/						2.5	-4.2	-1.1	-7.0	-1.8	-3.2	-4.3	-1.1	-5.7	-6.4
/ɛ/							1.9	-1.2	-1.5	0.2	-2.0	-7.0	-4.5	-6.1	-4.3
/ɜ/								3.9	-5.6	-1.5	-5.5	-1.6	-2.7	-1.3	-2.6
/eɪ/									2.5	-3.4	-2.7	-4.4	-4.3	-5.8	-6.5
/ɪ/										2.0	-0.9	-7.1	0.2	-2.2	-3.7
/i/											2.4	-4.4	-4.2	-5.8	-6.4
/oʊ/												2.8	-4.0	-2.5	-1.5
/ɔi/													4.9	0.1	-3.6
/ʊ/														2.6	-0.5
/u/															3.1

**Table 1.** Log-odds scoring matrix for stressed vowels. Each value represents the natural logarithm of the ratio of the likelihood of the pair matching in a rhyme versus the likelihood of the pair matching by chance. For example, the score of 1.6 for /ɔ/ and /a/ indicates that the pair is  $e^{1.6}$  (approximately five) times more likely to appear matched in a rhyme than by chance. Positive scoring pairs are highlighted.

For the example used in this section above, “breakin’ the mold” rhyming with “checkin’ my code,” the total score for the four syllable rhyme is 9.0: The first syllable scores  $-1.5(/eɪ/:/ɛ/)$  +  $2.6(/k/:/k/)$  +  $1.0(\text{matched strong stress}) = 2.1$ . The second syllable scores  $2.0(/ɪ/:/ɪ/)$  +  $2.2(/n/:/n/)$  +  $0.0(\text{matched weak stress}) = 4.2$ . The third syllable scores  $-1.4(/ə/:/ai/)$  +  $0.0(\text{weak stress}) = -1.4$ . The fourth syllable scores  $2.8(/oʊ/:/oʊ/)$  +  $(0.4+2.3)/(/ld/:/_d/)$  +  $1.0(\text{strong stress}) = 4.1$ .

	/b/	/tʃ/	/d/	/ð/	/f/	/g/	/dʒ/	/k/	/l/	/m/	/n/	/ŋ/	/p/	/r/	/s/	/ʃ/	/t/	/θ/	/v/	/z/	/ʒ/	*	*_
/b/	4.3	-4.8	1.1	0.4	-5.5	1.9	1.9	-6.9	-0.3	-0.5	-1.6	-5.5	0.1	-0.9	-1.6	-4.6	-1.0	-4.3	2.3	0.3	-2.5	-0.6	-1.5
/tʃ/		4.2	-1.6	-4.9	-0.3	0.2	0.4	1.5	-6.8	-6.6	-2.8	-5.5	1.1	-6.7	0.3	0.6	0.9	1.4	-6.1	-2.0	-2.5	-6.0	-2.6
/d/			2.3	-7.0	-7.6	0.1	0.2	-3.1	-1.7	-2.2	-2.2	-3.0	-1.8	-0.9	-9.0	-2.1	0.2	0.0	-0.2	0.0	-4.6	-0.2	1.2
/ð/				3.5	-5.6	-5.1	-4.2	-0.4	-0.2	-2.0	-7.5	-5.6	-6.2	-1.4	-7.0	-4.8	-0.3	1.3	2.8	1.1	-2.6	-6.0	-3.4
/f/					3.4	-1.2	-4.9	-0.3	-1.5	-1.3	-3.5	-1.6	1.1	-2.7	1.0	1.2	-0.9	4.0	0.6	-7.3	-3.2	-1.4	-2.9
/g/						4.1	1.8	0.0	-0.2	-1.0	-1.9	-5.7	-0.7	-0.8	-2.5	-4.9	-1.1	-4.5	0.3	-0.3	-2.7	-0.9	-2.8
/dʒ/							5.2	-6.3	-1.5	0.1	-0.5	-4.8	-0.2	-0.3	-0.6	0.6	-1.1	-3.6	1.4	1.0	4.1	-5.3	0.5
/k/								2.6	-2.9	-2.1	-2.6	-1.3	1.7	-2.1	-0.7	-0.6	0.9	0.5	-1.8	-3.1	-4.7	-0.9	-1.8
/l/									2.8	-1.8	-1.8	-2.8	-8.1	-0.5	-2.9	-6.6	-2.9	-6.3	-1.3	-1.6	-4.5	0.4	-1.0
/m/										2.7	1.8	0.7	-3.2	-1.2	-2.9	-1.1	-2.5	0.4	-0.6	-3.7	-4.2	-0.9	-1.7
/n/											2.2	1.2	-2.5	-1.0	-2.3	-0.7	-1.5	-0.6	-1.5	-2.1	-5.1	-0.4	-2.3
/ŋ/												4.1	-6.8	-2.7	-2.3	-5.3	-3.5	-5.0	-2.1	-2.0	-3.2	0.2	-3.9
/p/													3.3	-2.0	-1.1	-0.7	1.1	0.9	-0.5	-7.9	-3.8	-0.7	-0.8
/r/														2.8	-2.3	-0.8	-1.2	-6.1	-2.1	-2.2	-4.3	1.7	-0.7
/s/															2.6	2.4	-1.0	1.0	-2.4	0.5	0.0	0.5	0.6
/ʃ/																5.2	-0.6	-4.1	-1.3	-0.2	3.6	-5.8	-7.7
/t/																	1.7	1.6	-0.8	-9.2	-5.2	0.0	0.7
/θ/																		4.5	0.5	-6.1	-2.0	-5.4	-0.6
/v/																			2.9	-0.4	1.6	-1.2	-1.7
/z/																				2.6	3.0	-1.3	1.1
/ʒ/																					6.8	-3.7	-5.6

Table 2. Log-odds scoring matrix for consonants. Positive scoring pairs are highlighted.

### SCORING POTENTIAL RHYMES

With our probabilistic scoring method for matched syllables in place, we needed a procedure to identify internal and end rhymes. Our technique is a variant on local alignment (Durbin et al., 1999); for each syllable, we identified its closest preceding rhyming syllable, and longest preceding rhyming phrase within the current and previous lines. For example, given the line

*Unobtainable* to the **brain** it's *unexplainable* what the verse'll do (Monch, 1999)

from Pharoahe Monch's "Right Here," the middle "ain" syllables all rhyme, while the whole of "unexplainable" also rhymes with "unobtainable."

For every pair of consecutive lines in a set of lyrics, we first constructed a two-dimensional matrix of the score for every pair of syllables. Entries in this matrix (corresponding to pairs of syllables in the lines) were selected as "anchors" if they had scores above a threshold and contained a stressed syllable or were line-final. From these anchor positions, rhymes were extended forward, ensuring that the length-normalized score was above a syllable threshold. In addition to the iterative extension, a "jump"-type extension was also allowed, in which one or two syllables could be skipped over if the following syllable pair was an anchor type with score above a higher threshold. This was included since longer polysyllabic mosaic rhymes often contain one or two syllables that do not rhyme in the midst of three or four that do. A good example of this can be found in Fabolous' "Can't Deny It":

**I keep spittin', them clips copped on those calicos**  
**Keep shittin', with ziplocks of that Cali 'dro** (Fabolous, 2001)

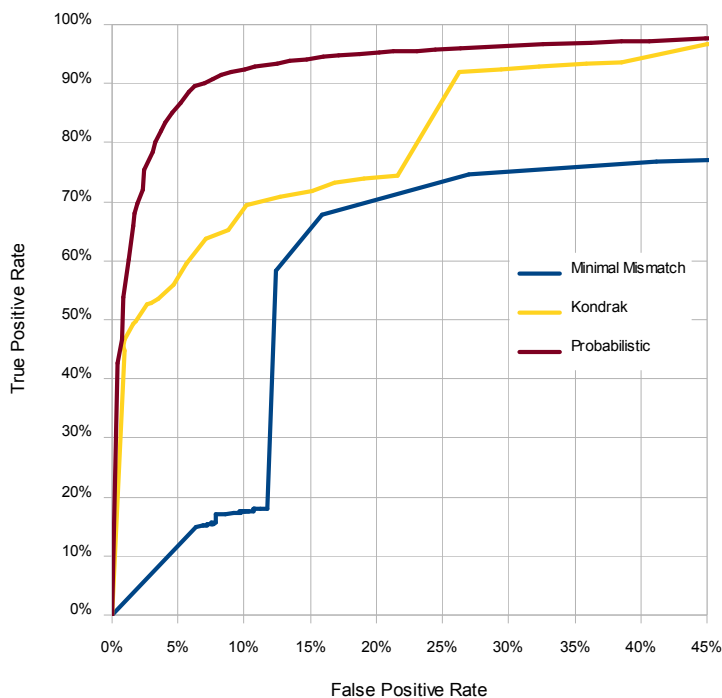
where the two lines rhyme in their entirety, with the exception of "them"/"with" and "those"/"that."

We filtered the set of rhymes to remove one-syllable rhymes including unstressed syllables, as these tended to be unintentional rhymes which happened to match by chance. After a set of rhymes was identified, we removed duplicates and consolidated consecutive and overlapping rhymes together.

### VALIDATING THE METHOD

Our first test verified that our probabilistic score for syllable rhyming was better at identifying perfect and imperfect rhymes than rules-based phonetic similarity measures. We did a 10-fold cross validation where we randomly selected 36 albums from the training data (consisting of 40 albums), trained a rhyme model for those albums, and used it to score the known rhyming lines from the other four albums. Lines scoring above the score threshold were true positives, while those scoring below the threshold were false negatives. We also used the model to score randomly selected lines from the remaining four albums (presumed to be non-rhymes). Randomly paired lines scoring above the threshold were considered to be false positives

(though they may, in fact, have been actual rhymes), while those scoring below the threshold were considered to be true negatives. We developed implementations of the minimal mismatch of articulatory features and Kondrak alignment (Kondrak, 2000) metrics to compare the performance of these scoring measures, which are based on the physical process of the human voice. We show receiver operator characteristic (ROC) curves comparing the true positive rate to false positive rate when varying the score threshold for each of the three methods in Figure 1. The probabilistic method significantly outperformed both simpler rules-based methods.



**Figure 1.** ROC curves for the three different scoring methods, comparing percentage of actual rhymes found on the y-axis with percentage of unrelated syllables detected as rhyming on the x-axis. The kink in the Minimal Mismatch curve is caused by a sharp decline in the number of detected rhymes when zero mismatched articulatory features are allowed, meaning only perfect rhymes are counted.

Next, we considered false positives and negatives for detected end rhymes, using the score threshold of 1.5 (meaning matched syllables are at least  $e^{1.5}$  (approximately 4.5) times more likely to rhyme than expected by chance). Out of 1000 pairs of unrelated random lines from our training data, 79 syllables were marked as parts of end rhymes (“false positives”) by our procedure. Of these, 22 were in fact true rhymes, with scores higher than 3.0. 30 were near-rhymes; that is, that they could be found (though less frequently) as line-final rhymes in actual lyrics. Usually scoring above 2.0, they included matches such as “stiff”/“fit,” “pen”/“thing,” and “cling”/“smothering,” with more than one articulatory difference or different stress. 14 matched-end syllables (often suffixes), typically with high scores (greater than 3.0). Examples such as “weaker”/“drummer” and “tappin”/“position,” may have exact matches, but are not relevant rhymes due to their lack of stress. The remaining 13 moderately high scoring (between 1.5 and 2.5) pairs featured either high consonant scores (like “bust”/“test”) or high vowel scores due to matching rare vowel sounds (“box”/“wrong”).

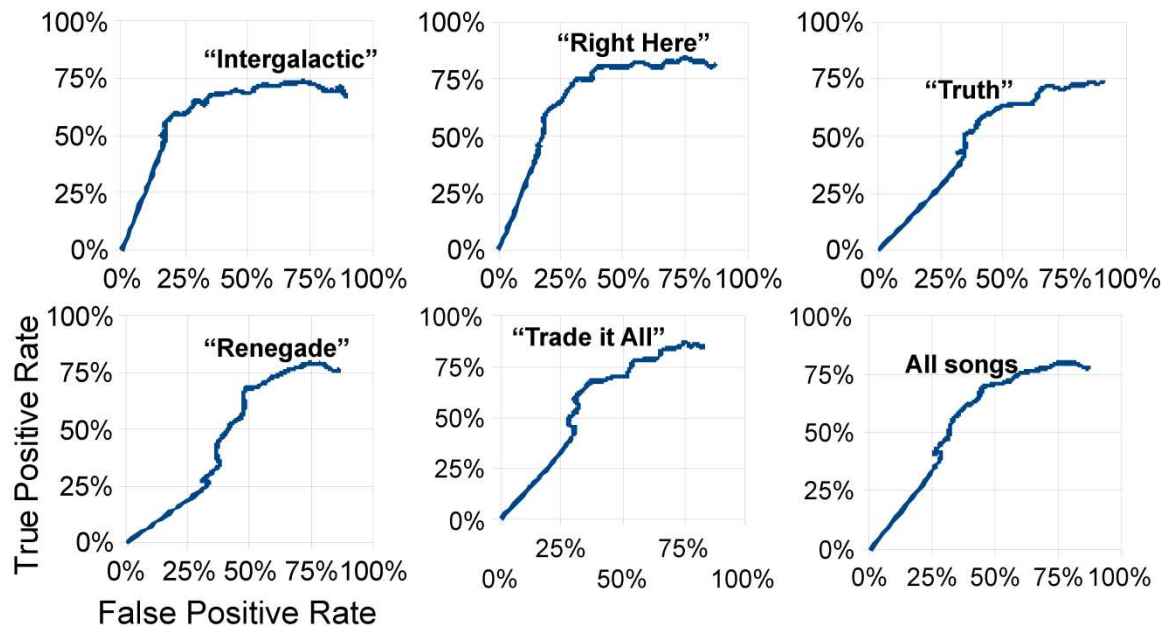
From a set of 1000 matched pairs of lines, we used the iterative method (moving backwards from the end of the line while scores for stressed syllables are positive) to see which true rhymes would be missed. Pairs with all such matches scoring less than 1.5 were marked and treated as false negatives. Out of 132 such syllables, 48 (the largest group) were moderately low scoring (between -1.0 and 1.5) pairs participating in polysyllabic and mosaic rhymes. A good example of this is “battery”/“battle me” in Eric B. and Rakim’s (1990) “No Omega”; many of these were flanked by high scoring pairs, and would be included in rhymes using the jump extension described in the above section. 35 were very low scoring pairs



(less than 0.0) which were either caused by words having been transcribed improperly in the dictionary or the lack of a true rhyme in the lyrics. 22 were caused by the rhyme start being extended too far back and starting with a low positive scoring pair. Again, this would not cause problems in our actual detection algorithm since, in that case, rhymes are extended forward from stressed anchors. 17 were caused by differences between the actual pronunciation and the dictionary's pronunciation ("poems" treated as one syllable, or "battles" specifically being pronounced to rhyme with "shadows"). Finally, 10 were caused by mismatched syllable stress due to deliberate modification of prosody in the performance.

The probabilistic model is quite good at finding both perfect and imperfect rhymes. Very few syllable pairs (less than 15 in the 1000 line pairs) scored highly without being perceivably rhyming, and most low scoring "true" rhyme pairs take part in complex mosaic and polysyllabic rhymes.

Finally, we used our model on a set of manually annotated rap lyrics, to measure the ability of the program to find both internal and line-final rhymes. We used five songs of varying styles: the Beastie Boys' "Intergalactic," a Grammy-winning song in the old-school style; Pharoahe Monch's "The Truth" (featuring Common and Talib Kweli) and "Right Here," which were annotated by Alim (2003) and feature high rhyme density and a complicated scheme; Jay-Z and Eminem's "Renegade," which features very high rhyme density; and Fabolous' "Trade It All (Part 2)," a song specifically mentioned by Alim for its prevalence of long (five or six syllable) rhymes. We show the ROC curves for this test set in Figure 2; the best overall performance is for specificity and sensitivity just above 60%.



**Figure 2.** Rhyme detection syllable ROC curves for test songs. The y-axis indicates the percentage of true rhymes identified by the algorithm, while the x-axis shows the percentage of automatically identified rhymes not considered to be true rhymes.

Most "false positives" are rhymes that were not annotated due to lack of rhythmic importance or accidental omission. False negatives included several where the performer created a rhyme from words that do not appear to rhyme as text, and some longer rhymes that were cut off prematurely due to too many non-rhyming syllables within them and lower scoring syllable pairs surrounding them. Finally, some rhymes were missed due to intervening rhymes being found between the rhyming parts, particularly when the threshold for rhymes was set low. This is especially evident in the ROC curves at lower cut-off thresholds, where true positive rates peak around 80% and begin to decline as the threshold is lowered.

As an illustration of the difficulty in finding the correct internal rhymes, consider the following lines from Eminem's first verse in "Renegade,"

*Usin' his music to steer it, sharin' his views and his merits*  
 But **there's** a huge interference, **they're** sayin you shouldn't **hear** it (Jay-Z, 2001),

in which he rhymes the five syllable phrase “music to steer it” with “views and his merits,” “huge interference,” and “you shouldn’t hear it,” along with numerous /*ɛr*/ sounds throughout. Since “music to” and “views and his” receive a low total similarity score (-0.6), the first five syllable rhyme is not detected, causing a false negative and a false positive since “views and his” is detected as rhyming with “usin’ his.” Similarly, “views and his” scores only 5.6 when matched with “huge inter-,” so that rhyme is not extended either, but “huge interference, they’re” scores 14.8 when matched with “music to steer it, shar[e],” resulting in another false positive.

### COMPARING RHYMING STYLE

We used our procedure to compute a variety of features about the rhymes in several sets of lyrics. These statistics include the number of syllables per line, the number of rhymes per line, the proportion of rhymes of different length, as well as the occurrences of the complex rhyming features (bridge, link, etc.) discussed earlier. The complete list of features calculated is detailed in Table 3.

Feature	Description
Syllables per Line	Average number of syllables per line
Syllables per Word	Average word length in syllables
Syllable Variation	Standard deviation of line lengths in syllables
Novel Word Proportion	Average percentage of words in the second line in a pair not appearing in the first
Rhymes per Line	Average number of detected rhymes per line
Rhymes per Syllable	Average number of detected rhymes per syllable
Rhyme Density	Total number of rhymed syllables divided by total number syllables
End Pairs per Line	Percentage of lines ending with a line-final rhyme
End Pairs Grown	Percentage of rhyming couplets in which the second line is more than 15% longer (in syllables) than the first
End Pairs Shrunk	Percentage of rhyming couplets in which the first line is more than 15% shorter (in syllables) than the second
End Pairs Even	Percentage of rhyming couplets neither grown or shrunk
Average End Score	Average similarity score of line final rhymes
Average End Syl Score	Average similarity score per syllable in line final rhymes
Singles per Rhyme	Percentage of rhymes being one syllable long
Doubles per Rhyme	Percentage of rhymes being two syllables long
Triples per Rhyme	Percentage of rhymes being three syllables long
Quads per Rhyme	Percentage of rhymes being four syllables long
Longs per Rhyme	Percentage of rhymes being longer than four syllables
Perfect Rhymes	Percentage of rhymes with identical vowels and codas
Line Internals per Line	Number of rhymes with both parts falling in the same line divided by total number of lines
Links per Line	Average number of link rhymes per line
Bridges per Line	Average number of bridge rhymes per line
Compounds per Line	Average number of compound rhymes per line
Chaining per Line	Total number of words or phrases involved in chain rhymes divided by total number of lines

**Table 3.** Description of higher-level rhyme features calculated.

### Genre Identification

We hypothesized that these features would show differences between genres of popular music, and calculated them for four sets of data: the top 10 songs from Billboard Magazine’s 2008 year-end Hot Rap Singles chart; the top 20 songs from the 2008 year-end Hot Modern Rock Songs chart; the first 400 lines of Milton’s (1667) *Paradise Lost* as a similar-sized, representative sample of a historically important form of non-rhyming verse (though we could have used a different sample for the baseline); and the top 10 songs

from the 1998 year-end Hot Rap Singles chart. To compare the verses most of all, the song lyrics were modified to remove intro/outro text, repeated lines, and additional choruses. Some of the interesting feature differences are shown in Table 4. High end rhyme scores are indicative of song lyrics in general (relative to unrhymed verse); rap has higher rhyme density, internal rhyme, link rhymes, and bridge rhymes. Interestingly, blank verse and rock lyrics have similar amounts of rhyming per line, but rock lyrics have more rhymes per syllable. Although *Paradise Lost* is written in iambic pentameter (meaning it should have exactly 10 syllables per line), its use of archaic words not found in the pronouncing dictionary and shifts in English pronunciation over time have it being detected as using more than 10 syllables per line. The data from 1998 and 2008 rap songs suggest that in their rhyming pattern, there has not been much shift in style, other than a possible increase in the amount of chain rhymes used.

Feature	Hot Rap 08	Hot Rap 98	Hot Rock	Blank Verse
Number of Lines	476	613	502	400
Number of Syllables	4646	6492	4053	4146
Syllables per Line	9.76	10.59	8.07	10.37
Number of Rhymes	794	1118	476	393
Rhymes per Line	1.67	1.82	0.95	0.98
Rhymes per Syllable	0.17	0.17	0.12	0.09
Rhyme Density	0.28	0.27	0.19	0.12
Average Total End Score	5.28	5.21	4.36	2.49
Average End Syllable Score	3.75	3.67	4.01	2.28
Singles per Rhyme	0.64	0.60	0.72	0.78
Doubles per Rhyme	0.23	0.29	0.15	0.18
Triples per Rhyme	0.08	0.06	0.04	0.03
Quads per Rhyme	0.02	0.03	0.05	0.00
Longs per Rhyme	0.03	0.02	0.04	0.01
Internals per Line	0.62	0.60	0.27	0.28
Links per Line	0.20	0.28	0.13	0.16
Bridges per Line	0.43	0.48	0.28	0.40
Chaining per Line	0.32	0.18	0.15	0.07

**Table 4.** Comparison of selected rhyme features for different genres. Some of the more interesting differences are highlighted.

### Use of Rhyme Features within Hip Hop

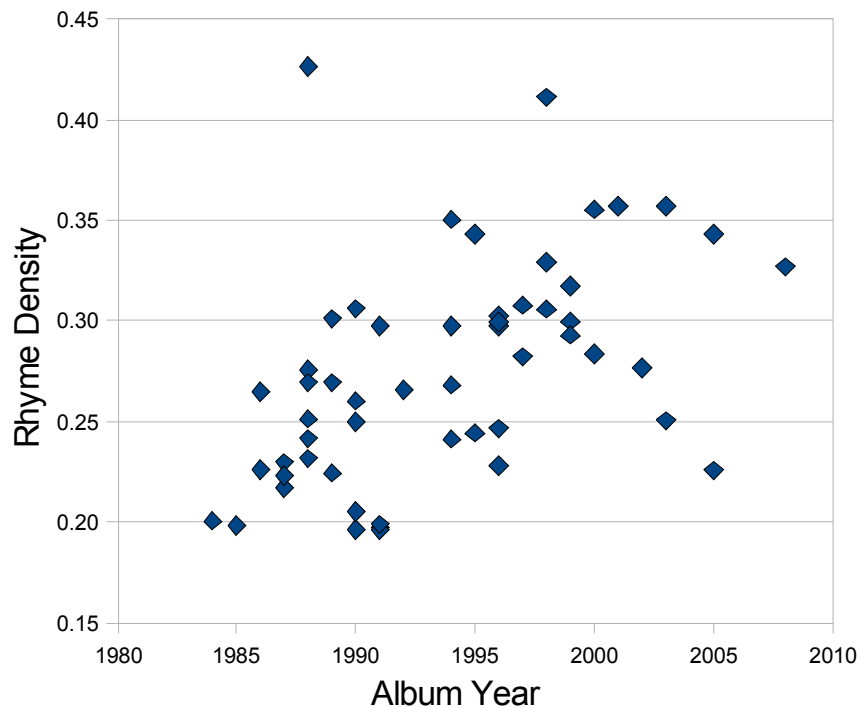
We also hypothesized that features of individual rappers might be informative, so we produced these statistics for popular albums by 25 famous MCs from a diverse range of styles and eras. These include some of the best-selling rappers (RIAA, 2010), as well as those considered by many to be among the best of all time (MTV, 2006) (Adaso, 2007): Run-DMC, LL Cool J, the Beastie Boys, Rakim, KRS-One of Boogie Down Productions, Chuck D of Public Enemy, Big Daddy Kane, Slick Rick, Kool G Rap, Ice Cube, MC Hammer, Scarface, Redman, Nas, Andre 3000 of Outkast, The Notorious B.I.G., 2Pac, Bone Thugs-n-Harmony, Jay-Z, DMX, Eminem, Nelly, Fabolous, 50 Cent, and Lil' Wayne. Since many of the statistics involve the position of rhymes in relation to the end and middle of lines, we listened to each album to ensure that lyrics were transcribed such that each line of text corresponded to a single bar or measure of music in the song. We again focused on the rapped verses, removing any lyrics which were either spoken, sung, or performed by guest artists. The results indicate that many of these features can be quite characteristic of different artists' styles.

For example, early rappers Run-DMC (1984, 1986), LL Cool J, (1985, 1987), and the Beastie Boys' (1986, 1989) old-school style uses less rhyme with around 1.7 rhymes per line and a rhyme density of 0.22 compared to the overall average among artists of 2.0 rhymes per line and a rhyme density of 0.27. Rakim (Eric B. & Rakim, 1987, 1988), recording around the same time but known for his more complex style, is detected as using more triplet rhymes (9%) than previous artists (4%). Later Golden Age rappers,

such as KRS-One (Boogie Down Productions, 1987, 1988), Big Daddy Kane (1988, 1990), and Kool G Rap (1989, 1990), display even higher rhyme density scores, as well as a tendency to move away from line final rhymes. This is especially the case with Chuck D (Public Enemy, 1988, 1990), whose 0.32 end rhymes per line is amongst the lowest in our collection and quite a bit lower than the average (0.40). Rival rappers Notorious B.I.G. (1994) and Tupac Shakur's (1995) early albums display some fairly similar style characteristics: about two rhymes per line, 30% of rhymes being two syllables long, and 9% being longer. However, Biggie's lines are shorter in length, with, on average, 11.1 syllables to 2Pac's 12.4.

Faster rappers like Andre 3000 (Outkast, 1994, 1996, 1998) and Bone Thugs-n-Harmony (1995) can squeeze in the most syllables per line (14.7 and 17.2 compared to the average of 11.6), allowing them to achieve the most rhymes per line (3.2 and 3.6) since they rap more words which can be matched with others in adjacent lines. Andre 3000 can also be considered to be one of the most eloquent MCs, using about 1.4 syllables per word, compared to the artist average of 1.25. DMX (1998, 1999), on the other hand, uses the shortest words of any rapper (around 1.19 syllables), though he does tend to use more mosaic rhymes, with 17% of his rhymes being longer than two syllables. Artists from the early 2000s like Eminem (1999, 2000), and especially Fabolous (2001, 2003) also favour longer rhymes, with 19% and 30% respectively of their rhymes being longer than two syllables. The most recent MC in the group, Lil' Wayne (2005, 2008) manages one of the highest rhyme density scores (0.33) while using some of the shortest lines (11.1 syllables) since the early '90s. The full set of feature data by album is included in Appendix 1.

These data lead us to the observation of a few key trends in the development of rhyming style in hip hop over time. Most significant is the increase in rhyme density as MCs began to use more rhymes and longer rhymes. Rhyme density and year have a moderate correlation with an  $r$ -value of 0.48 ( $p < 0.001$ ). This is displayed in Figure 3. Low rhyme density scores at 1990 and 1991 correspond to albums by Ice Cube (1990, 1991) and Scarface (1991) (whose raps are more story-oriented and feature less intricate rhyming) and M.C. Hammer (1990, 1991) (who generally performed dance-style rap songs). The low scores at 2003 and 2005 are for 50 Cent (2003, 2005), whose removal from the data set would result in a correlation with  $r$ -value 0.57.



**Figure 3.** An illustration of the increase in rhyme density (number of rhymed syllables divided by total number of syllables) over time. Outliers at the top-left and bottom-right of the graph are Run-D.M.C.'s *Tougher Than Leather* and 50 Cent's first two albums, respectively.

The increase in longer rhymes can also be illustrated by the declining usage of one-syllable rhymes over time (see Figure 4). Year has a negative correlation with the percentage of rhymes being one syllable long with an  $r$ -value of  $-0.64$  ( $p < 0.001$ ) (and matching positive correlations with the proportion of triples, quads, and longer rhymes, with  $r$ -values all greater than or equal to  $0.5$  and  $p$ -values less than  $0.001$ ).

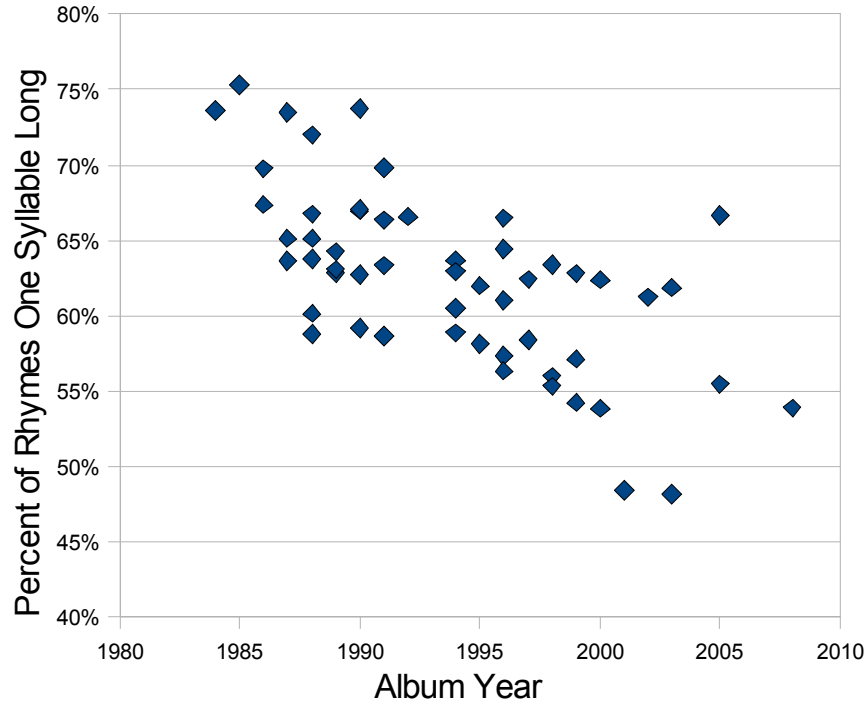
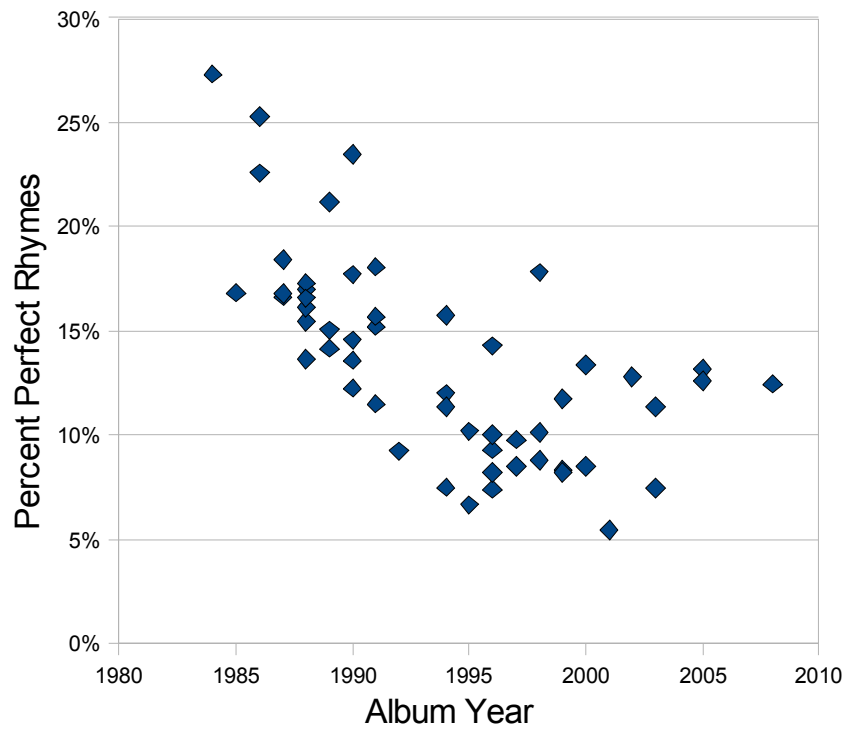
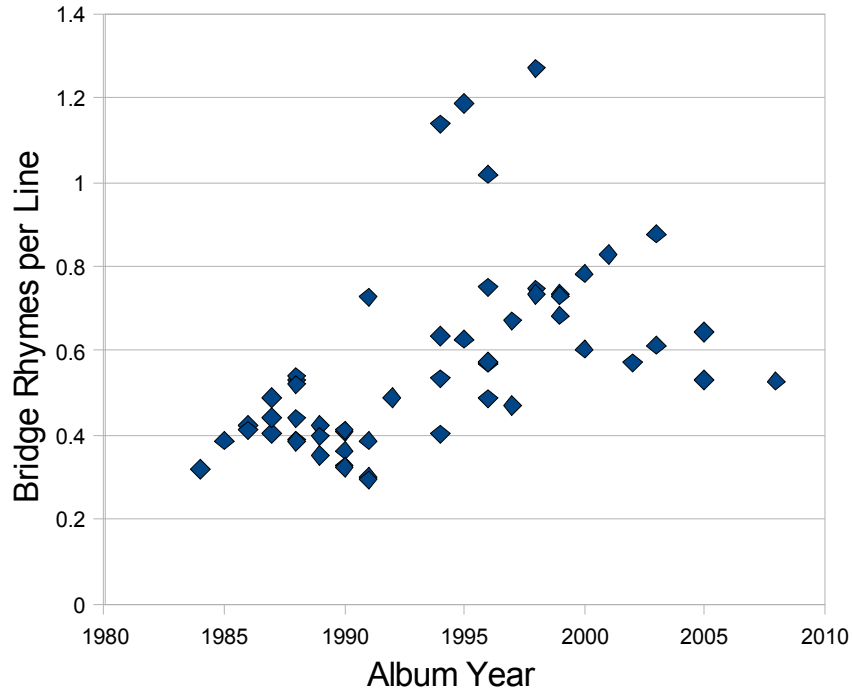


Figure 4. An illustration of the decreasing usage of monosyllabic rhymes over time.



**Figure 5.** An illustration of the decreasing use of perfect rhymes (in which both vowels and consonant codas match exactly) over time.

Another interesting phenomenon is the increasing use of imperfect rhymes, resulting in year having negative correlations with average end syllable score ( $r = -0.75, p < 0.001$ ) and percent of perfect rhymes ( $r = -0.63, p < 0.001$ ). See Figure 5 for an illustration of this relationship. Finally, we see increases in the usage of the more complex features, such as link rhymes per line, which has a correlation of 0.42 ( $p = 0.001$ ) with year, and bridge rhymes per line, which has a correlation of 0.53 ( $p < 0.001$ ) with year (see Figure 6).



**Figure 6.** An illustration of the increasing use of bridge rhymes (containing internal words/phrases in consecutive lines) over time.

### CLASSIFYING ARTISTS USING RHYME FEATURES

After observing that most of the MCs in our collection displayed fairly consistent styles between albums, which often tended to be quite distinctive, we hypothesized that we might be able to classify lyrics by rapper, using only these statistical features. We broke our data set into “songs,” which we treated as segments of at least 40 lines (corresponding to, at the minimum, two 16 line verses and 8 lines of chorus), and calculated the features for each of these. This resulted in 603 songs over the 53 albums. We fed this set of instances into the Weka Data Mining Software (Hall et al., 2009) and classified the data using a Simple Logistic Regression. We used 10-fold cross-validation, in which a model was trained on 90% of the instances and used to classify the remaining 10%. The results were surprisingly good, with 314 (52%) of the instances being classified correctly.

The full classification results produced a weighted F-measure of 0.516. The F-measure of a class is a balanced combination of the precision ( $p$ ) and recall ( $r$ ) for that class; the precision is the percentage of instances assigned by the model to the class which actually belong to the class; the recall is the percentage of instances actually belonging to the class which are assigned to it by the model. The formula is  $F_1 = 2 * p * r / (p + r)$ . All rappers were most often classified as themselves, meaning that that their songs were identified as being by the correct artist more often than they were identified as being by any other particular artist, with the exception of KRS-One who was most often classified as Rakim. He, along with Jay-Z, Slick

Rick, and Nelly, were among the most difficult to classify using the regression model, with F-measures all falling below 0.3.

Considering we used fewer than 25 fairly simple statistical features and no semantic information, the classification results are much higher than the 4% correctness we would expect by chance. A purely random classifier would achieve our level of accuracy with probability well under  $10^{(-50)}$ . An obvious comparison for classification would be to use a standard bag-of-words model, which performs much better at identifying rappers. Using a naïve Bayes bag-of-words classifier and the same 10-fold cross-validation as above, 552 (91.5%) instances were classified correctly with a weighted F-measure of 0.91. However, this is not a very informative method for this genre. Rappers have a very strong inclination to name-drop in their lyrics, including the use of their own names, nicknames, and record label and group names. This can be seen in the attributes of the naïve Bayes model, as the highest weighted ones for the majority of rappers were usually one of these names. If a song has the word “jigga” in it, it is very likely to be by Jay-Z; seeing “weezy” is evidence of a Lil’ Wayne song; “slim” and “shady” are indicative of Eminem; and Scarface really likes to say “Brad” (his real first name).

Furthermore, the classification errors made by a bag-of-words model tell us nothing about the style of the MCs in question. When Nelly was identified as Jay-Z, it was because his vocabulary emphasizes words like “dough,” “ice,” and “game,” not because he rhymes like Jay-Z. Andre 3000 was often misclassified as Lil’ Wayne as they share a Southern vernacular, but their rapping styles are dissimilar. Conversely, classification errors made using the rhyme features can raise very interesting comparisons of rhyming style. For example, Run-D.M.C. were often confused with the Beastie Boys. This error was to be expected due to Run-D.M.C.’s influence upon the Beastie Boys and their association with Def Jam Recordings. In fact, Run-D.M.C. wrote or co-wrote some of the songs from the Beastie Boys’ *Licensed to Ill* (1986), including “Paul Revere” and “Slow and Low,” which was originally a Run-D.M.C. recording. According to Beastie Boy Ad-Rock, “our sound right then was desperately trying to sound like Run-D.M.C.” (Ro, 2005).

Other artists misclassified as each other included Ice Cube and Scarface, who we perceive to have a similar story-telling style, light in the use of intricate rhyming patterns, and often using uneven line lengths. The Notorious B.I.G. was most often confused with one of his influences Kool G Rap, and hip hop scholar Adam Bradley finds the similarity connection between superstars Eminem and Jay-Z to be the most interesting (Bradley, personal communication, 2010). The full classification results confusion matrix is displayed in Table 5.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y		
3	6	0	2	0	0	0	2	0	0	0	0	3	0	0	0	0	0	0	1	0	0	1	0	0	0	←classified as	
1	14	1	0	0	2	0	1	3	0	0	1	0	0	0	1	0	0	0	1	0	1	0	1	0	0	a = KRS-One	
0	0	11	4	0	0	1	1	0	0	3	0	0	0	0	3	0	0	2	1	0	0	0	0	0	0	b = Rakim	
1	1	3	15	0	1	0	1	0	0	1	0	0	0	2	0	0	1	1	1	1	3	0	0	0	0	c = Nas	
0	1	0	0	17	0	0	0	0	1	0	0	1	0	0	4	0	0	0	0	0	1	0	0	1	0	d = 2Pac	
0	1	0	0	0	16	0	0	1	0	0	1	2	0	1	2	0	2	0	2	0	3	1	0	0	0	e = Run-D.M.C.	
0	0	1	0	0	1	17	0	0	0	2	0	0	1	1	1	0	1	1	0	0	0	1	0	0	0	f = 50 Cent	
1	2	0	0	0	0	0	11	1	1	6	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	g = Lil’ Wayne
0	4	0	0	1	0	0	3	10	0	1	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	h = Kool G Rap
0	0	0	0	1	0	1	1	1	14	2	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	i = Big Daddy Kane
1	0	2	1	0	0	2	2	1	0	20	0	0	0	0	2	0	0	1	0	0	1	0	0	0	0	0	j = Chuck D
0	0	0	0	0	2	0	0	1	2	0	13	0	6	0	0	2	0	0	0	0	0	0	0	0	0	1	k = The Notorious B.I.G.
1	0	0	0	1	0	1	2	0	0	1	11	1	0	0	0	2	1	0	0	2	1	0	1	0	0	1	l = Ice Cube
0	0	0	0	3	1	0	0	0	0	4	1	10	0	0	0	0	0	0	0	0	0	0	1	0	0	0	m = LL Cool J
0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	1	0	2	0	0	0	0	2	0	1	0	0	n = Scarface
0	1	1	5	0	0	3	0	0	0	2	0	0	0	1	6	0	5	0	2	0	2	4	0	0	0	0	o = Fabolous
0	1	0	0	4	0	0	0	2	0	0	1	0	0	0	12	0	1	0	0	0	0	0	0	0	0	0	p = Jay-Z
0	1	3	0	0	1	3	0	0	0	0	0	0	0	3	0	15	0	0	0	0	2	0	0	0	0	0	q = Beastie Boys
0	1	0	2	1	1	0	0	0	0	3	2	2	0	0	1	1	5	0	1	0	1	0	1	0	0	0	r = Eminem
0	0	2	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	14	0	0	2	0	0	0	0	s = Slick Rick
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	12	0	0	0	0	0	t = Redman
0	1	0	1	0	4	1	0	0	0	0	0	0	0	3	0	1	0	1	0	5	1	0	1	0	1	0	u = Andre 3000
0	0	0	0	0	2	1	0	0	1	1	0	0	0	2	1	0	1	0	0	1	17	0	0	0	0	0	v = Nelly
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	13	0	0	w = DMX
0	0	0	1	0	2	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	1	1	0	0	10	0	x = Bone Thugs-n-Harmony
																										y = MC Hammer	

**Table 5.** Confusion matrix for rappers using a simple logistic regression with statistical rhyme features. Each row corresponds to songs by the artist indicated at the right, while the columns display the frequency

that songs were associated with particular rappers. The class(es) under which each rapper is most often classified is highlighted.

The rhymes we detect and the features we calculate are especially indicative of rhyming style in rap music, and they allow us to build surprisingly useful statistical characterizations of different MCs. That these features are indicative particularly of rapping style is further supported by their relative weakness at characterizing other types of music. We performed a similar classification experiment using ten top-selling dance/pop artists popular around the same time as the MCs in our collection: Michael Jackson, ABBA, Celine Dion, Madonna, the Backstreet Boys, Cher, Janet Jackson, Mariah Carey, Britney Spears, and Eurythmics. A Simple Logistic Regression classifier using rhyme features correctly identified only 62 of 234 instances (26.5%) with an F-measure of 0.26, which is only marginally better than the 10% classification accuracy we would expect by chance.

## APPLICATIONS AND DISCUSSION

### Style Modification

Given that we have a reasonably accurate statistical characterization of various rappers' rhyming styles, we can begin to consider other applications for which rhyming style can be analyzed. One such application would be in inferring an artist's stylistic consistency from the accuracy by which they can be classified. For the most easily identified rappers in our collection, we can identify a single distinctive characteristic which explains their statistical uniqueness. For Andre 3000 and Bone Thugs-n-Harmony, it is their speedy flow resulting in relatively more syllables per line. For Fabolous, it is his extensive usage of extremely long multisyllabic rhymes. For other relatively well classified artists such as Run-D.M.C., Chuck D, and Redman, it is less obvious what makes their rhyming style so consistently distinct. Conversely, artists who are poorly classified can be considered to have a multitude of rhyme styles with which they can "switch up" their flow. In the song "22 Two's" from his debut album *Reasonable Doubt*, Jay-Z boasts that "I don't follow any guidelines cause too many niggas ride mine/ so I change styles every two rhymes" (Jay-Z, 1996), and his resistance to classification can be seen as evidence of this constant changing of styles.

When speaking about MCs' diversity of style, it should be noted that three of the artists in our collection are in fact not single rappers, but groups: Run-D.M.C., the Beastie Boys, and Bone Thugs-n-Harmony. However, this does not have much of an effect on the results for two main reasons. In the case of Run-D.M.C. and the Beastie Boys, many of their raps are delivered in a way such that individual members perform alternating lines (sometimes even trading parts of lines), and certain phrases (especially line-final rhymes in the case of the Beastie Boys) are performed by the whole group. This makes it extremely difficult to separate one member's contribution to the song from another's. Bone Thugs-n-Harmony generally has each of its members performing a shortened "mini-verse" of eight to twelve consecutive lines as part of longer verses. However, each of the five rappers in the group rhyme similarly enough to each other and differently enough from other rappers in the collection, that the group as a whole can be said to have a distinct style.

Rappers also sometimes consciously modify their style, perhaps in a deliberate attempt to imitate another artist's rhyme technique. A well-known example of this phenomenon including artists from our collection occurs in the song "Notorious Thugs" from The Notorious B.I.G.'s *Life After Death* (1997), which features Bone Thugs-n-Harmony as guest performers. Classifying Biggie's verse in the song using the rhyme feature logistic regression described above "incorrectly" identifies him as Bone Thugs-n-Harmony. However, his rap in this song was in fact deliberately performed so as to mimic the Bone Thugs' style. According to producer Steven "Stevie J." Jordan, "after Bone Thugs went in there and ripped it, Big took it home for a minute. He was like, 'I ain't laying mine. I got to wait. This style ain't what I'm used to.'" As Bone Thugs member Layzie Bone put it, "When Biggie did our style, that's when Bone received respect for our shit" (XXL Magazine, 2003).

### Ghostwriter Identification

Another interesting application of statistical rhyme style characterization is in the detection of ghostwriting in hip hop songs. While the term can refer to a range of practices in the industry, in this



domain it generally refers to raps written by an artist other than the performer. We performed a small experiment on identifying ghostwriters using artists from our collection known to write for other rappers. Ice Cube was known as the primary lyricist for seminal gangsta rap group N.W.A. and is credited with writing most of their songs. We classified Eazy-E's "Boyz-n-the-Hood" and N.W.A.'s "Express Yourself" (performed by Dr. Dre), both written by Ice Cube (Eazy-E, 1987; N.W.A., 1988), using the rhyme feature logistic regression classifier and "Boyz-n-the-Hood" was in fact identified as Ice Cube. Nas famously wrote for Will Smith on his multi-platinum solo album *Big Willie Style* (Smith, 1997), but using the classifier on songs "Just Cruisin'" and "Gettin' Jiggy Wit It" did not identify Nas as the rapper. However, songs written by Jay-Z for producers Dr. Dre ("Still D.R.E."), (Dr. Dre, 1999) and Timbaland ("Indian Carpet") (Timbaland & Magoo, 2001) are both identified by the classifier as Jay-Z. As a comparison, using the naïve Bayes bag-of-words classifier on these songs only correctly identifies one of them: "Boyz-n-the-Hood" is classified as Ice Cube, likely due to the inclusion of Cube-indicative words such as "fools," "hoe," and "nappy," all of which are in the top 10 attributes for his class. The results are detailed in Table 6.

Song Title	Writer	Performer	Rhyme Classification	Bag-of-Words Classification
Boyz-n-the-Hood	Ice Cube	Eazy-E	Ice Cube	Ice Cube
Express Yourself	Ice Cube	Dr. Dre	Big Daddy Kane	Big Daddy Kane
Just Cruisin'	Nas	Will Smith	The Notorious B.I.G	Big Daddy Kane
Gettin' Jiggy Wit It	Nas	Will Smith	Jay-Z	Jay-Z
Still D.R.E.	Jay-Z	Dr. Dre	Jay-Z	Fabulous
Indian Carpet	Jay-Z	Timbaland	Jay-Z	MC Hammer

**Table 6.** Classification results for ghostwritten songs using rhyme and bag-of-words features. Correctly identified writers are highlighted.

It should be noted that the songs we used here all had the "ghostwriter" included in the credits and also include the performer as a writer, meaning that they are in fact co-written. However, for the songs performed by producers (not known primarily for their rhyming abilities), we may assume that the co-writing MC had a much greater contribution to the lyrics. This may explain why the classifier does not identify Nas in the songs co-written by Nas and Will Smith, an established rapper himself. Perhaps the collaboration of the two writers resulted in a conglomerate style not characteristic of either rapper. The ability to classify Jay-Z and Ice Cube in the other cases does suggest that we may be able to identify ghostwriters even when they are not credited, which may often be the case. For example, even though rapper Skillz is perhaps best known as being a writer for other rappers (XXL Magazine, 2007), his ASCAP entry only has one song in which he does not rap himself (ASCAP, 2010). (Ironically, this is Will Smith's "Lost & Found," in which he asks, "Why should I try to flow the way ya'll flow?" (Smith, 2005).) Developing a statistical profile of Skillz's rhyming style may allow us to identify songs to which he has made uncredited contributions.

## Content-based Recommendation

Finally, with our set of rhyme style features, we can make larger-scale comparisons between different rappers allowing for content-based recommendation in hip hop. Using normalized Euclidean distance in the 24-dimensional feature space, we built a "phylogeny" of our albums using the Neighbor-Joining algorithm (Saitou & Nei, 1987) (Figure 7). Even with this simple distance metric, the artists cluster in a reasonable way. Most of the artists tend to fall in small clusters with their own albums and albums by similar artists. The largest distinction is between old-school style rap (generally produced before the 1990s) and newer hip hop, though even within the older artists, the more intricate rhymers (Rakim and Big Daddy Kane) are branched off. Among the newer artists, there is a split between the less rhyme-dense mid-'90s (2Pac, Redman, and the Notorious B.I.G.) and other performers, who are further subdivided into the faster (Andre 3000 and Bone Thugs) and slower rappers. We can also see that the most difficult artists to classify

(Jay-Z, KRS-One, Slick Rick, but surprisingly not Nelly) generally have albums that are *not* very similar to their other albums, indicating their diversity or progression in style.

With an embedding of MCs in a high-dimensional rhyme style space such as this one, we could easily find the two or three rappers most similar to any other artist (given enough of their lyrics to calculate the rhyme statistics). This could allow lesser-known performers to promote themselves by highlighting their similarity to more famous rappers, or let music recommendation systems make suggestions based on the rhyming styles their users prefer. Suggesting artists based on their rhyming style would be an important step towards true content-based recommendation in hip hop, a genre in which the majority of the musical information is in the lyrics.

## CONCLUSIONS

Using a probabilistic scoring model, we were able to identify both perfect and imperfect rhymes with a higher level of accuracy than simpler rules-based methods. The heuristic rhyme detection methods achieved moderate success at finding both internal and line-final rhymes in song lyrics. More importantly, statistical features of these rhymes did correspond to real world characterizations of rhyme style, and many are quite consistent within individual artists' lyrics and varied between different artists. Using these features we illustrate trends in hip hop over time, including the increase in rhyme density and complex rhyme features, as well as the decrease in perfect rhymes and single syllable rhymes. Finally, these features can be used to classify artists with remarkably high accuracy, resulting in possible applications for style comparison and ghostwriting detection.

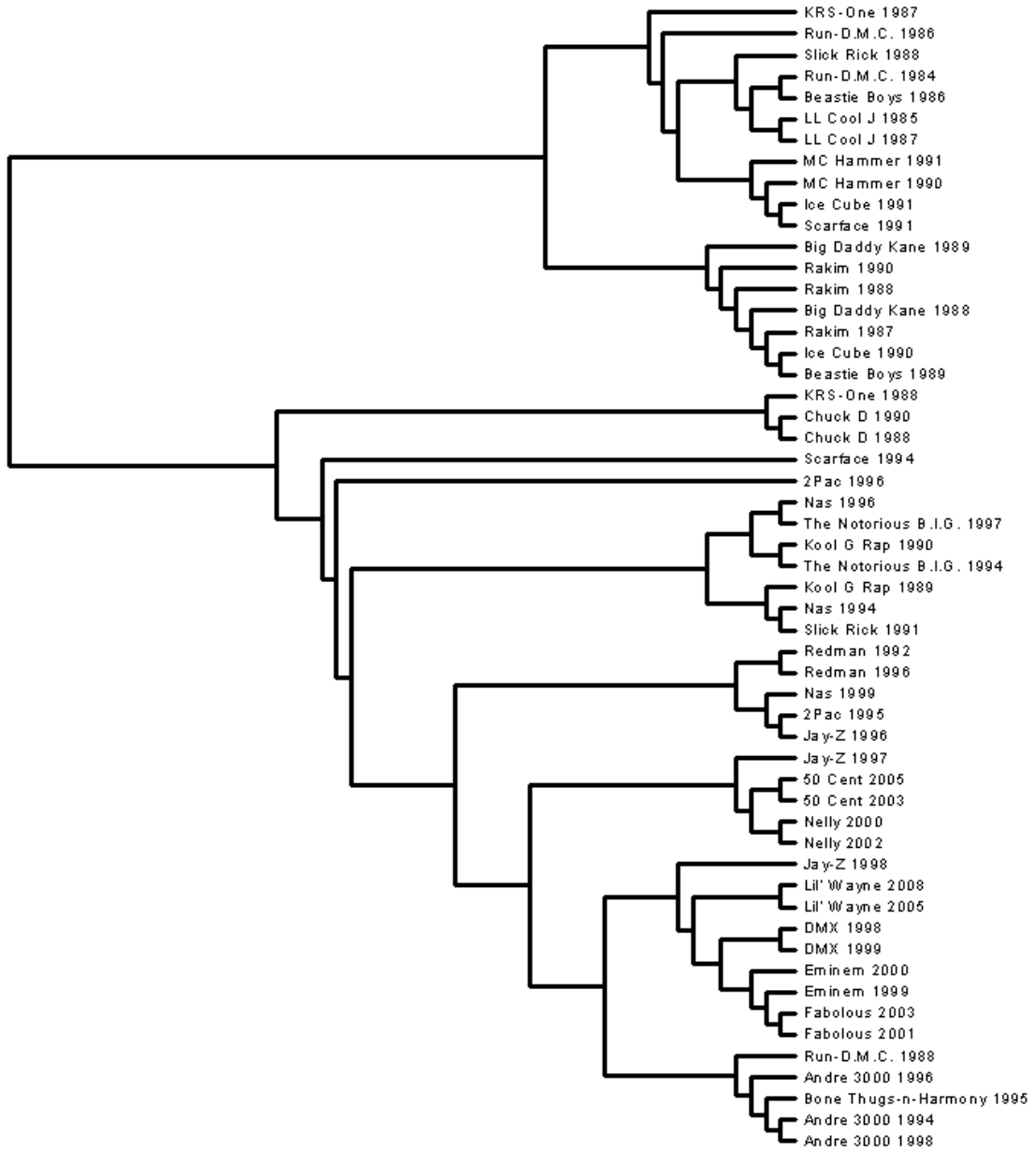


Figure 7. A hierarchical clustering of the albums built by the Neighbor-Joining algorithm.

## REFERENCES

- Adaso, H. (2007). *50 Greatest MCs of Our Time (1987 – 2007)*. About.com: Rap/Hip-Hop. <http://rap.about.com/od/toppicks/ss/Top50Emcees.htm>
- Alim, H. S. (2003). On some serious next millennium rap ish: Pharoahe Monch, hip hop poetics, and the internal rhymes of internal affairs. *Journal of English Linguistics*, Vol. 31, No. 1, pp. 60–84.
- American Society of Composers, Authors, and Publishers (ASCAP). (2010). *Works written by Lewis, Donnie Ray*. <http://ascap.com/ace/search.cfm?mode=search>
- Bradley, A. (2009). *Book of Rhymes: The Poetics of Hip Hop*. New York: Basic Civitas Books.
- Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1999). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press.
- Elovitz, H.S., Johnson, R.W., McHugh, A., Shore, J.E. (1976). Automatic translation of English text to phonetics by means of letter-to-sound rules. *Interim Report Naval Research Lab*. Washington, DC. <http://www.speech.cs.cmu.edu/comp.speech/Section5/Synth/text.phoneme.3.html>
- Fujihara, H., Goto, M., Ogata, J. (2008). Hyperlinking lyrics: a method for creating hyperlinks between phrases in song lyrics. *Proceedings of the International Conference on Music Information Retrieval*, pp. 281–286.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, Vol. 11, No. 1, pp. 10-18.
- Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 89, No. 22, pp. 10915–10919.
- Hirjee, H. & Brown, D.G. (2009). Automatic detection of internal and imperfect rhymes in rap lyrics. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pp. 711–716.
- Holtman, A. (1996) *A Generative Theory of Rhyme: An Optimality Approach*. Ph.D. dissertation, Utrecht Institute of Linguistics.
- Kawahara, S. (2007). Half rhymes in Japanese rap lyrics and knowledge of similarity. *Journal of East Asian Linguistics*, Vol. 16, No. 2, pp. 113–144.
- Kleedorfer, F., Knees, P., Pohle, T. (2008) Oh oh oh whoah! Towards automatic topic detection in song lyrics. *Proceedings of the International Conference on Music Information Retrieval*, pp. 287–292.
- Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 288–295.
- Lenzo, K. (2007). *The CMU Pronouncing Dictionary*. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
- Mayer, R., Neumayer, R., Rauber, A. (2008). Rhyme and style features for musical genre classification by song lyrics. *Proceedings of the International Conference on Music Information Retrieval*, pp. 337–342.
- Milton, J. (1667). *Paradise Lost*. London: Samuel Simmons.
- MTV. (2006). *The Greatest MCs of All Time*. [http://www.mtv.com/bands/h/hip\\_hop\\_week/2006/emcees/](http://www.mtv.com/bands/h/hip_hop_week/2006/emcees/)
- Ogbar, J. O. G. (1999). Slouching toward Bork: The Culture Wars and Self-Criticism in Hip-Hop Music. *Journal of Black Studies*, Vol. 30, No. 2, pp. 164-183.

The Recording Industry Association of America (RIAA). (2010). *Top Selling Artists*.  
<http://www.riaa.com/goldandplatinumdata.php?table=tblTopArt>

Ro, R. (2005). *Raising Hell: The Reign, Ruin, and Redemption of Run-D.M.C. and Jam Master Jay*. New York: HarperCollins Publishers.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, Vol. 4, No. 4, pp. 406–425.

Wei, B., Zhang, C., Ogiwara, M. (2007). Keyword generation for lyrics. *Proceedings of the International Conference on Music Information Retrieval*, pp. 121-122.

XXL Magazine. (2003). *The Making of Life After Death: Many Men*.  
<http://www.xxlmag.com/online/?p=405>

XXL Magazine. (2007). *Skillz: I'm The Dopest*. <http://www.xxlmag.com/online/?p=12639>

## DISCOGRAPHY

- Beastie Boys. (1986). *Licensed to Ill*, Def Jam Recordings.
- Beastie Boys. (1989). *Paul's Boutique*, Capitol Records.
- Big Daddy Kane. (1988). *Long Live the Kane*, Cold Chillin' Records.
- Big Daddy Kane. (1990). *It's a Big Daddy Thing*, Cold Chillin' Records.
- Bone Thugs-n-Harmony. (1995). *E. 1999 Eternal*, Ruthless Records.
- Boogie Down Productions. (1987). *Criminal Minded*, B-Boy Records.
- Boogie Down Productions. (1988). *By All Means Necessary*, Jive Records.
- Dr. Dre. (1999). *2001*. Aftermath Entertainment.
- DMX. (1998). *It's Dark and Hell is Hot*, Ruff Ryders Entertainment.
- DMX. (1999). *...And Then There Was X*, Ruff Ryders Entertainment.
- Eazy-E. (1987). *Boyz-N-The-Hood*. Ruthless Records.
- Eminem. (1999). *The Slim Shady LP*, Aftermath Entertainment.
- Eminem. (2000). *The Marshall Mathers LP*, Aftermath Entertainment.
- Eric B. and Rakim. (1987). *Paid in Full*, 4<sup>th</sup> and B'way Records.
- Eric B. and Rakim. (1988). *Follow the Leader*, Uni Records.
- Eric B. and Rakim. (1990). *Let the Rhythm Hit 'Em*, MCA Records.
- Fabulous. (2001). *Ghetto Fabulous*, Elektra Records.
- Fabulous. (2003). *Street Dreams*, Desert Storm Records.
- Ice Cube. (1990). *AmeriKKKa's Most Wanted*. Priority Records.
- Ice Cube. (1991). *Death Certificate*. Priority Records.
- Jay-Z. (1996). *Reasonable Doubt*. Roc-A-Fella Records.
- Jay-Z. (2001). *The Blueprint*, Roc-A-Fella Records.
- Kool G Rap and DJ Polo. (1989). *Road to the Riches*, Cold Chillin' Records.
- Kool G Rap and DJ Polo. (1990). *Wanted: Dead or Alive*. Cold Chillin' Records.
- Lil' Wayne. (2005). *Tha Carter II*, Cash Money Records.
- Lil' Wayne. (2008). *Tha Carter III*, Cash Money Records.
- LL Cool J. (1985). *Radio*, Def Jam Recordings.

- LL Cool J. (1987). *Bigger and Deffer*, Def Jam Recordings.
- MC Hammer. (1990). *Please Hammer, Don't Hurt 'Em*. Capitol Records.
- MC Hammer. (1991). *Too Legit to Quit*. Capitol Records.
- The Notorious B.I.G. (1994). *Ready to Die*, Bad Boy Records.
- The Notorious B.I.G. (1997). *Life After Death*. Bad Boy Records.
- N.W.A. (1988). *Straight Outta Compton*. Ruthless Records.
- Outkast. (1994). *Southernplayalisticadillacmuzik*, LaFace Records.
- Outkast. (1996). *ATLiens*, LaFace Records.
- Outkast. (1998). *Aquemini*, LaFace Records.
- Pharoahe Monch. (1999). *Internal Affairs*, Rawkus Records.
- Public Enemy. (1988). *It Takes a Nation of Millions to Hold Us Back*. Def Jam Recordings.
- Public Enemy. (1990). *Fear of a Black Planet*, Def Jam Recordings.
- Run-D.M.C. (1984). *Run-D.M.C.*, Profile Records.
- Run-D.M.C. (1986). *Raising Hell*, Profile Records.
- Scarface of the Geto Boys. (1991). *Mr. Scarface is Back*. Rap-a-Lot Records.
- Scarface. (1994). *The Diary*. Rap-a-Lot Records.
- Timbaland and Magoo. (2001). *Indecent Proposal*. Blackground Records.
- Will Smith. (1997). *Big Willie Style*. Columbia Records.
- Will Smith. (2005). *Lost and Found*. Interscope Records.
- 2Pac. (1995). *Me Against the World*, Jive Records.
- 50 Cent. (2003). *Get Rich or Die Tryin'*. Aftermath Entertainment.
- 50 Cent. (2005). *The Massacre*. Aftermath Entertainment.

## APPENDIX

## Appendix 1. Full list of rhyme features for albums in the collection.

Artist	Year	Syllables per Line	Syllables per Word	Syllable Variation	Novel Word Proportion	Rhymes per Line	Rhymes per Syllable	Rhyme Density	End Pairs per Line	End Pairs Grown	End Pairs Shrunk	End Pairs Even	Average End Score	Average End Syl Score
Run-D.M.C.	1984	10.40	1.21	1.82	0.91	1.48	0.14	0.20	0.54	0.30	0.17	0.53	5.51	4.59
Run-D.M.C.	1986	10.88	1.21	2.05	0.91	1.96	0.18	0.26	0.47	0.26	0.18	0.56	5.58	4.49
Run-D.M.C.	1988	10.17	1.21	2.21	0.89	2.51	0.25	0.42	0.43	0.21	0.18	0.62	6.18	4.27
LL Cool J	1985	11.49	1.29	2.03	0.92	1.68	0.15	0.20	0.46	0.27	0.18	0.55	5.28	4.28
LL Cool J	1987	11.71	1.23	1.89	0.92	1.87	0.16	0.21	0.47	0.37	0.16	0.48	5.26	4.58
Beastie Boys	1986	10.81	1.21	1.98	0.88	1.63	0.15	0.22	0.50	0.24	0.20	0.56	5.99	4.68
Beastie Boys	1989	11.30	1.27	2.38	0.90	1.67	0.15	0.22	0.47	0.29	0.25	0.46	6.57	4.52
Rakim	1987	11.18	1.24	1.93	0.90	1.64	0.15	0.23	0.51	0.23	0.29	0.48	6.10	4.07
Rakim	1988	10.83	1.32	2.29	0.92	1.85	0.17	0.27	0.46	0.25	0.32	0.43	6.10	3.96
Rakim	1990	10.62	1.27	1.94	0.92	1.76	0.17	0.25	0.44	0.21	0.25	0.54	6.00	3.85
KRS-One	1987	11.77	1.30	2.14	0.92	1.74	0.15	0.22	0.42	0.18	0.19	0.63	5.88	4.10
KRS-One	1988	11.20	1.30	2.72	0.91	1.90	0.17	0.26	0.43	0.26	0.28	0.46	5.46	3.87
Chuck D	1988	10.35	1.26	2.42	0.90	1.75	0.17	0.25	0.33	0.28	0.23	0.49	5.46	4.00
Chuck D	1990	9.78	1.25	2.43	0.92	1.66	0.17	0.24	0.31	0.27	0.30	0.43	5.40	3.89
Big Daddy Kane	1988	10.76	1.29	1.90	0.92	1.71	0.16	0.24	0.44	0.29	0.19	0.52	6.33	4.13
Big Daddy Kane	1989	10.46	1.25	2.03	0.93	1.82	0.17	0.26	0.42	0.31	0.23	0.46	6.04	4.01
Slick Rick	1988	11.40	1.20	2.11	0.91	1.86	0.16	0.23	0.48	0.36	0.16	0.48	5.83	4.40
Slick Rick	1991	12.62	1.26	2.65	0.91	2.25	0.18	0.29	0.34	0.28	0.12	0.59	6.40	3.80
Kool G Rap	1989	10.94	1.32	2.04	0.93	2.13	0.19	0.29	0.34	0.29	0.21	0.50	6.49	4.25
Kool G Rap	1990	10.96	1.29	2.47	0.92	2.08	0.19	0.29	0.35	0.34	0.30	0.36	6.33	4.08
Ice Cube	1990	10.04	1.20	2.21	0.91	1.35	0.13	0.19	0.44	0.31	0.29	0.40	6.04	4.16
Ice Cube	1991	9.65	1.24	2.58	0.93	1.30	0.13	0.19	0.48	0.49	0.22	0.29	6.08	4.01
MC Hammer	1990	8.93	1.21	3.13	0.92	1.30	0.15	0.20	0.46	0.51	0.13	0.36	4.74	4.14
MC Hammer	1991	10.94	1.21	2.57	0.91	1.51	0.14	0.19	0.45	0.37	0.17	0.46	5.13	4.16
Scarface	1991	10.37	1.22	2.90	0.90	1.37	0.13	0.20	0.50	0.55	0.22	0.24	5.69	3.92
Scarface	1994	11.81	1.22	2.87	0.88	1.69	0.14	0.24	0.47	0.45	0.19	0.36	6.35	3.87
Redman	1992	11.69	1.20	2.21	0.91	2.11	0.18	0.26	0.34	0.26	0.21	0.53	5.01	3.41
Redman	1996	11.33	1.27	2.02	0.93	1.89	0.17	0.24	0.37	0.32	0.19	0.49	5.09	3.51
Nas	1994	12.45	1.27	2.42	0.92	2.32	0.19	0.29	0.33	0.31	0.18	0.50	5.91	3.77
Nas	1996	11.84	1.30	2.62	0.92	2.17	0.18	0.29	0.38	0.30	0.34	0.36	5.36	3.44
Nas	1999	12.37	1.26	2.05	0.91	2.35	0.19	0.29	0.36	0.27	0.24	0.49	5.03	3.46
Andre 3000	1994	14.58	1.37	3.25	0.92	3.28	0.22	0.34	0.32	0.41	0.21	0.38	5.05	3.68
Andre 3000	1996	14.68	1.35	3.31	0.91	2.78	0.19	0.29	0.31	0.30	0.19	0.51	5.44	3.84
Andre 3000	1998	14.76	1.48	5.02	0.91	3.56	0.24	0.40	0.29	0.46	0.24	0.30	4.93	3.41
The Notorious B.I.G.	1994	11.08	1.26	2.31	0.92	1.94	0.17	0.26	0.31	0.38	0.25	0.36	5.80	3.90
The Notorious B.I.G.	1997	10.62	1.25	2.52	0.92	2.00	0.19	0.30	0.35	0.26	0.29	0.44	5.34	3.65
2Pac	1995	12.35	1.30	2.50	0.91	1.96	0.16	0.24	0.32	0.25	0.24	0.51	5.08	3.58
2Pac	1996	12.46	1.29	2.61	0.92	1.85	0.15	0.22	0.38	0.26	0.27	0.46	5.69	3.74
Bone Thugs-n-Harmony	1995	17.23	1.24	3.22	0.89	3.58	0.21	0.33	0.21	0.27	0.27	0.47	4.59	3.47
Jay-Z	1996	12.08	1.28	2.39	0.92	2.15	0.18	0.29	0.28	0.30	0.18	0.52	5.09	3.36
Jay-Z	1997	12.13	1.22	2.23	0.90	2.11	0.17	0.27	0.42	0.22	0.19	0.59	5.47	3.63
Jay-Z	1998	11.64	1.22	2.32	0.89	2.18	0.19	0.32	0.35	0.24	0.27	0.49	5.51	3.64
DMX	1998	12.30	1.18	2.44	0.88	2.06	0.17	0.30	0.33	0.28	0.16	0.55	6.34	3.83
DMX	1999	12.32	1.19	2.24	0.87	2.11	0.17	0.29	0.33	0.21	0.23	0.57	6.05	3.88
Eminem	1999	12.41	1.24	2.50	0.88	2.17	0.17	0.31	0.44	0.29	0.22	0.49	6.47	3.78
Eminem	2000	12.39	1.23	2.55	0.88	2.40	0.19	0.34	0.43	0.24	0.18	0.58	5.40	3.20
Nelly	2000	11.58	1.23	2.30	0.89	2.09	0.18	0.28	0.43	0.23	0.14	0.63	4.95	3.74
Nelly	2002	11.81	1.23	2.55	0.88	1.95	0.17	0.27	0.40	0.24	0.17	0.59	4.98	3.70
Fabulous	2001	11.45	1.21	2.24	0.89	1.86	0.16	0.35	0.34	0.25	0.32	0.43	6.33	3.41
Fabulous	2003	11.47	1.24	2.42	0.87	1.92	0.17	0.35	0.32	0.25	0.24	0.51	6.10	3.41
50 Cent	2003	11.73	1.20	2.48	0.88	1.82	0.16	0.25	0.44	0.21	0.18	0.61	5.96	3.85
50 Cent	2005	11.57	1.20	2.71	0.90	1.75	0.15	0.22	0.45	0.19	0.20	0.61	5.27	3.87
Lil' Wayne	2005	11.45	1.27	2.66	0.88	2.21	0.19	0.33	0.40	0.32	0.26	0.42	5.30	3.33
Lil' Wayne	2008	10.71	1.24	2.92	0.87	1.92	0.18	0.32	0.44	0.33	0.30	0.38	5.43	3.59



Artist	Year	Singles per Rhyme	Doubles per Rhyme	Triples per Rhyme	Quads per Rhyme	Longs per Rhyme	Perfect Rhymes	Line Internals per Line	Links per Line	Bridges per Line	Compounds per Line	Chaining per Line
Run-D.M.C.	1984	74%	20%	4%	1%	2%	27%	0.47	0.16	0.32	0.06	0.21
Run-D.M.C.	1986	67%	26%	4%	2%	1%	22%	0.88	0.19	0.42	0.12	0.23
Run-D.M.C.	1988	59%	24%	10%	4%	3%	17%	1.30	0.27	0.52	0.11	0.41
LL Cool J	1985	75%	18%	4%	1%	1%	16%	0.61	0.22	0.38	0.10	0.18
LL Cool J	1987	74%	21%	4%	1%	0%	18%	0.73	0.19	0.48	0.11	0.25
Beastie Boys	1986	70%	21%	4%	4%	1%	25%	0.55	0.18	0.40	0.07	0.19
Beastie Boys	1989	64%	27%	6%	2%	1%	21%	0.61	0.18	0.41	0.09	0.15
Rakim	1987	63%	24%	9%	2%	1%	17%	0.61	0.13	0.39	0.08	0.13
Rakim	1988	59%	28%	9%	3%	1%	13%	0.73	0.22	0.44	0.09	0.19
Rakim	1990	62%	28%	7%	2%	1%	11%	0.68	0.24	0.40	0.08	0.12
KRS-One	1987	65%	28%	5%	2%	1%	16%	0.62	0.27	0.43	0.10	0.12
KRS-One	1988	65%	25%	6%	3%	2%	17%	0.67	0.27	0.53	0.11	0.22
Chuck D	1988	67%	24%	6%	2%	0%	15%	0.79	0.25	0.38	0.10	0.20
Chuck D	1990	67%	26%	6%	2%	0%	13%	0.79	0.21	0.36	0.11	0.27
Big Daddy Kane	1988	63%	27%	7%	2%	1%	16%	0.71	0.18	0.38	0.07	0.19
Big Daddy Kane	1989	62%	28%	8%	2%	1%	14%	0.80	0.25	0.35	0.10	0.26
Slick Rick	1988	72%	21%	4%	2%	1%	16%	0.63	0.24	0.51	0.11	0.22
Slick Rick	1991	58%	29%	8%	3%	2%	10%	0.89	0.32	0.71	0.17	0.18
Kool G Rap	1989	62%	29%	6%	2%	0%	13%	1.04	0.37	0.38	0.15	0.25
Kool G Rap	1990	58%	33%	7%	2%	1%	13%	1.01	0.32	0.40	0.14	0.22
Ice Cube	1990	67%	25%	5%	1%	1%	18%	0.40	0.19	0.32	0.04	0.13
Ice Cube	1991	67%	26%	5%	2%	1%	15%	0.34	0.18	0.30	0.05	0.12
MC Hammer	1990	74%	18%	5%	2%	1%	23%	0.41	0.12	0.32	0.09	0.17
MC Hammer	1991	70%	24%	4%	2%	0%	18%	0.56	0.21	0.29	0.05	0.20
Scarface	1991	63%	30%	5%	2%	1%	15%	0.33	0.16	0.38	0.05	0.09
Scarface	1994	59%	28%	7%	3%	3%	12%	0.47	0.22	0.53	0.07	0.14
Redman	1992	66%	27%	4%	2%	1%	9%	1.02	0.27	0.48	0.17	0.28
Redman	1996	66%	27%	5%	1%	0%	9%	0.74	0.30	0.48	0.11	0.18
Nas	1994	60%	30%	7%	3%	1%	7%	1.02	0.35	0.63	0.25	0.18
Nas	1996	57%	33%	6%	3%	1%	7%	0.90	0.34	0.56	0.16	0.17
Nas	1999	63%	27%	5%	4%	1%	8%	0.90	0.38	0.71	0.20	0.15
Andre 3000	1994	64%	27%	5%	2%	1%	16%	1.49	0.39	1.07	0.57	0.43
Andre 3000	1996	65%	26%	6%	3%	1%	14%	1.23	0.27	0.98	0.41	0.29
Andre 3000	1998	64%	22%	7%	5%	3%	18%	1.71	0.35	1.22	0.73	0.51
The Notorious B.I.G.	1994	62%	29%	6%	2%	1%	10%	0.83	0.39	0.41	0.13	0.18
The Notorious B.I.G.	1997	58%	30%	8%	3%	1%	9%	0.82	0.37	0.46	0.12	0.27
2Pac	1995	61%	30%	7%	2%	0%	10%	0.69	0.33	0.62	0.14	0.12
2Pac	1996	60%	31%	6%	2%	0%	9%	0.63	0.27	0.57	0.12	0.12
Bone Thugs-n-Harmony	1995	58%	30%	7%	3%	1%	6%	1.85	0.37	1.14	0.65	0.76
Jay-Z	1996	56%	31%	9%	3%	1%	8%	0.83	0.31	0.73	0.16	0.17
Jay-Z	1997	62%	24%	9%	4%	1%	8%	0.74	0.31	0.65	0.12	0.22
Jay-Z	1998	56%	28%	8%	6%	2%	10%	0.80	0.31	0.73	0.13	0.23
DMX	1998	55%	27%	10%	6%	3%	8%	0.70	0.30	0.72	0.14	0.20
DMX	1999	57%	28%	9%	4%	2%	11%	0.74	0.32	0.72	0.11	0.14
Eminem	1999	54%	27%	12%	5%	2%	7%	0.77	0.30	0.67	0.15	0.21
Eminem	2000	54%	28%	10%	6%	3%	8%	0.85	0.36	0.75	0.16	0.18
Nelly	2000	62%	28%	8%	2%	1%	13%	0.74	0.34	0.59	0.14	0.18
Nelly	2002	61%	27%	7%	3%	2%	12%	0.72	0.27	0.55	0.12	0.21
Fabulous	2001	49%	19%	14%	12%	7%	5%	0.50	0.22	0.81	0.06	0.13
Fabulous	2003	48%	23%	12%	9%	8%	7%	0.55	0.18	0.86	0.09	0.16
50 Cent	2003	62%	26%	7%	3%	2%	11%	0.57	0.21	0.61	0.10	0.16
50 Cent	2005	66%	25%	5%	2%	1%	13%	0.56	0.22	0.52	0.10	0.12
Lil' Wayne	2005	55%	29%	9%	4%	3%	12%	0.83	0.34	0.63	0.12	0.29
Lil' Wayne	2008	53%	30%	9%	5%	3%	12%	0.68	0.29	0.52	0.11	0.27